Intégrales floues pour la reconnaissance de la parole

Julie Mauclair, Laurent Wendling, David Janiszek

Intégrales floues pour la reconnaissance de la parole

Julie Mauclair Laurent Wendling David Janiszek

LIPADE - Université Paris Descartes
45 rue des Saints Pères, 75006 PARIS
{julie.mauclair,laurent.wendling,david.janiszek}@parisdescartes.fr

Résumé

Cet article présente des travaux sur l'aggrégation de mesures de confiance en reconnaissance de la parole en utilisant des techniques provenant de la logique floue. Contrairement aux approches précédentes de la littérature utilisant principalement la notion de probabilité, nous nous proposons d'observer la notion d'incertitude des hypothèses de reconnaissance et la notion de possibilité grâce à la logique floue. Quatre mesures de confiance provenant chacune d'une partie différente d'un système de reconnaissance ont été développées. Plusieurs méthodes d'aggrégation sont étudiées afin d'en évaluer la capacité à améliorer les performances des mesures de confiance prises séparement. Ces méthodes sont évaluées en termes de Confidence Error Rate (CER) sur un corpus d'émissions radiophoniques en Français. Des méthodes probabilistes sont ainsi comparées à des techniques d'aggrégation floues, parmi lesquelles l'intégrale de Choquet parvient à améliorer les performances de manière significative en termes de CER.

Mots Clef

Reconnaissance de la parole, Logique Floue, Aggrégation de mesures de confiance

Abstract

This paper presents a study on merging confidence measures using fuzzy logic. Instead of the previous approaches using the notion of probability, we propose to observe the uncertainty of the recognition hypotheses and the notion of possibility thanks to fuzzy reasoning. Four different confidence measures are developed, coming from different parts of a speech recognizer. Various merging methods are studied to improve the performance of the confidence measures. The methods are evaluated in terms of Confidence Error Rate (CER) on a French broadcast news corpus. They are compared to some fuzzy logic aggregation techniques among which the technique based on the Choquet Integral yields to a significant improvement in terms of CER.

Keywords

Speech Recognition, Fuzzy Logic, Aggregation of Confidence Measures

1 Introduction

Les mesures de confiance ont été étudiées à de nombreuses reprises pour des applications en traitement de la parole [15]. À l'aide d'un seuil, un score de confiance associé à une hypothèse de reconnaissance peut par exemple, permettre l'acceptation ou le rejet de cette hypothèse [1]. Les mesures de confiance peuvent également être utilisées pour sélectionner les hypothèses issues de traitements automatiques d'enregistrements audio bruts pour accroître le corpus d'apprentissage des modèles acoustiques [19, 29]. Elles peuvent aussi guider la stratégie de confirmation de systèmes de dialogue homme-machine ou détecter les mots hors-vocabulaire [25] et lors d'applications d'identification de la langue [22].

En reconnaissance de la parole, les mesures de confiance donnent une estimation sur la justesse d'une hypothèse. Pour fournir de telles estimations, diverses parties d'un système de reconnaissance sont utiles et donnent une indication sur la fiabilité et la pertinence du système.

Pour bénéficier des qualités de chacune des mesures, il s'agit de trouver une technique d'aggrégation performante. Le résultat attendu de cette agrégation est d'obtenir une décision finale plus robuste et qui va améliorer la capacité de classification d'une mesure seule. Les opérateurs les plus utilisés pour cette combinaison sont : le minimum, maximum, moyenne arithmétique, moyenne géométrique, produit ou encore la moyenne quadratique. D'autres techniques comme les Machines à Vecteurs de Support (SVM), les réseaux de neurones ou les Modèles de Mixture de Gaussiennes (GMM) sont utilisées pour combiner les scores de confiance [30].

Cependant, aucune de ces familles d'opérateurs ne prend en compte les interactions possibles entre les différents éléments de l'aggrégation. L'intégrale de Choquet est alors considérée car elle permet d'utiliser ces interactions pour généraliser divers opérateurs d'aggrégation en choisissant des mesures floues spécifiques comme la moyenne

arithmétique pondérée, la moyenne ordonnée pondérée, les statistiques d'ordre n et la médiane [11, 12, 18]. Les intégrales floues, et l'intégrale de Choquet en particulier, ont été utilisées avec succès en tant qu'opérateurs d'aggrégation dans plusieurs applications de reconnaissances des formes [10] et de reconnaissance de la parole [20].

Le but de cet article est alors d'explorer la manière dont les intégrales floues et la sélection de mesures peuvent améliorer les scores de confiance associés à des hypothèses de reconnaissance en traitement de la parole.
Premièrement, nous présenteront les différentes mesures de confiance développées pour notre étude. Le troisième paragraphe décrira les différentes techniques d'aggrégation utilisées en tant que références. Dans le paragraphe 4, les deux intégrales floues utilisées vont être introduites. Ensuite, nous détaillerons le corpus utilisé ainsi que les métriques d'évaluation. Des résultats sur chacune des techniques utilisées seront également décrits. Enfin, nous présenterons comment la sélection de mesures présentes dans l'aggrégation peut permettre d'améliorer des résultats.

2 Mesures de confiance

Soit un ensemble de N mots émis par le système de reconnaissance $\{w_1, \ldots, w_N\}$. Chaque mot w est alors associé à une mesure de confiance $m(w)$ qui respecte les propriétés suivantes :

Propriété 1 : La mesure de confiance est située dans l'intervalle $[0, 1]$;
Propriété 2 : La mesure de confiance est une bonne approximation de la probabilité qu'un mot w soit correct.

En conséquence de la deuxième propriété, la moyenne arithmétique de la mesure de confiance sur le nombre de mots doit être une approximation du taux de reconnaissance correcte sur les mots émis (on ne compte pas les suppressions), que l'on notera $t_{correct}$ ($\frac{1}{N}\sum_{i=1}^{N} m(w_i) \approx t_{correct}$).

2.1 Mesure Acoustique (AC)

Cette mesure de confiance est calculée à partir des scores de vraisemblance du modèle acoustique contraint par l'arbre lexical et par le modèle de langage utilisés lors du processus de reconnaissance ainsi que des scores du modèle acoustique utilisé sans contraintes (boucle de phonèmes) [19] :

$$m^*_{ac}(w) = \frac{1}{N_f(w)}[\log P(Y|\lambda_C) - \log P(Y|\lambda_L)] \quad (1)$$

où w est le mot reconnu sur N_f frames, Y est la séquence d'observations acoustiques, $P(Y|\lambda_C)$ est le score acoustique donné par les modèles du système, et $P(Y|\lambda_L)$ est le score acoustique donné par une boucle de phonème non contrainte.
Afin que cette mesure respecte les propriétés précédentes et soit considérée comme une mesure de confiance, nous proposons une normalisation qui permettra de rester dans l'intervalle $[0, 1]$ grâce à une transformation de type sigmoidal [3]. Cette transformation est présentée dans la formule suivante :

$$m_{AC}(w) = \frac{\exp\left(\frac{m^*_{AC}-\mu}{\sigma}\right) + a}{\exp\left(\frac{m^*_{AC}-\mu}{\sigma}\right) + 1} \quad (2)$$

où μ and σ sont respectivement la moyenne et l'écart-type des mesures acoustiques initiales sur les mots d'un corpus de développement. Pour s'approcher de la deuxième propriété définie au début du paragraphe et qui indique que la moyenne d'une mesure de confiance idéale doit approcher le taux de mots émis bien reconnus $t_{correct}$, nous nous servons du même corpus de développement et obtenons pour a :

$$a = 2 * t_{correct} - 1$$

Cette mesure sera notée $m_{AC}(w)$.

2.2 Mesure basée sur le repli du modèle de langage (LMBB)

Cette mesure a été proposée dans [19]. À partir du modèle de langage utilisé pour la reconnaissance, cette mesure prend en compte l'ordre du n-gramme le plus élevé qui peut être associé au mot visé par la mesure de confiance et à l'historique de taille $n-1$ de ce mot.
Par exemple, si la séquence de mots «... il est temps de ...» est reconnue en utilisant un modèle de langage quadrigramme et que le quadrigramme [il est temps de] a été observé dans le corpus d'apprentissage du modèle de langage, alors le mot 'de' sera associé à l'ordre 4. Par contre, si ce quadrigramme n'a pas été observé, mais que le trigramme [est temps de] l'a été, alors le mot 'de' sera associé à l'ordre 3. De la même manière, ce mot pourrait être associé à l'ordre 2 ou à l'ordre 1 le cas échéant, et même à l'ordre 0 dans le cas peu courant où les mots hors-vocabulaire peuvent être traités. Afin de ne pas distinguer un nombre de classes différentes trop important qui seraient difficiles à bien modéliser sans grande quantité de données d'apprentissage, nous ne prendrons pas les valeurs réelles des ordres associés aux mots voisins mais leur position relative par rapport à l'ordre associé au mot visé : plus grand (+), plus petit (-) ou égal (=). Ceci permet de réduire le nombre de classes possibles.
Pour illustrer ce propos, prenons par exemple la séquence de mots «... il est temps de lire ce livre...» et supposons :
– que le quadrigramme [il est temps de] et le trigramme [est temps de] n'ont pas été observés dans le corpus d'apprentissage du modèle de langage, alors que le bigramme [temps de] l'a été : le mot 'de' est associé à l'ordre 2 ;
– que le quadrigramme [est temps de lire] n'a pas été observé dans le corpus d'apprentissage alors que le trigramme [temps de lire] l'a été : le mot 'lire' est associé à l'ordre 3 ;

– que le quadrigramme [temps de lire ce] a été observé dans le corpus d'apprentissage : le mot 'ce' est associé à l'ordre 4.

Dès lors, la classe de comportement du mot 'lire' sera associé à l'étiquette (-,3,+), puisque le mot 'lire' est associé à l'ordre 3, son voisin de gauche est associé à un ordre inférieur (-) de valeur 2 et son voisin de droite est associé à un ordre supérieur (+) de valeur 4.

En comparant un ensemble de transcriptions automatiques dont les mots sont marqués par ce type d'étiquettes, et en ayant pour les enregistrements audio de ces transcriptions des transcriptions manuelles, il est aisé de calculer le taux d'erreur de reconnaissance pour les mots qui composent chacune de ces classes. Ce taux d'erreur est le rapport entre le nombre de mots $n_{err}(cl)$ mal reconnus (substitutions ou insertions) contenus dans une classe cl sur le nombre de mots $n_{mots}(cl)$ qui composent cette classe (pour un ensemble de transcriptions donné). Ainsi, pour un mot w associé à la classe cl, la valeur $m_{lmbb}(w)$ donnée par la mesure de confiance LMBB se calcule à partir d'un corpus d'apprentissage composées de transcriptions automatiques et manuelles par la formule suivante :

$$m_{LMBB}(w) = 1 - \frac{n_{err}(cl)}{n_{mots}(cl)} \quad (3)$$

Il a été montré dans [19] qu'il existe une réelle corrélation entre les triplets et le taux d'erreur associés aux mots.

La mesure de confiance associée au comportement du repli du modèle de langage est notée $m_{LMBB}(w)$.

2.3 Probabilité *a posteriori*

Les probabilités *a posteriori* peuvent être calculées à partir des listes des N meilleures hypothèses [8], ou encore à partir des graphes de mots [7, 15]. En fait, la probabilité *a posteriori* d'un mot est le ratio entre la probabilité d'un mot et la somme des probabilités de toutes les autres hypothèses alternatives. Ces probabilités sont une combinaison des scores fournis par les modèles acoustiques et linguistique.

Dans les listes des N meilleures hypothèses la probabilité *a posteriori* d'un mot est le rapport de la somme des probabilités des occurrences de ce mot à une position donnée parmi les N hypothèses, sur la somme de toutes les probabilités des mots situés à la même position, incluant celles des occurrences du mot courant.

Dans les approches basées sur les graphes de mots ou les réseaux de confusions, la probabilité *a posteriori* est la généralisation de l'approche précédente où la segmentation en mots et la profondeur de l'espace de recherche sont mieux pris en considération.

Ici, nous utilisons un réseau de confusion basé sur la technique utilisée dans [17] pour calculer les probabilités *a posteriori* des mots. Cette technique permet de regrouper dans le graphe de mots tous les arcs associés au même mot avec différentes prononciations et les mots qui sont en concurrence. La mesure de confiance basée sur les probabilités *a posteriori* des mots sera notée $m_{WP}(w)$.

2.4 Mesure basée sur une transformation linéaire par morceaux de la probabilité *a posteriori*

Les probabilités *a posteriori* des mots issues du réseau de confusion peuvent directement être utilisée en tant que mesure de confiance. Cependant, il apparaît que la probabilité *a posteriori* d'un mot tende à sous-estimer la pertinence réelle de ce mot [7]. Cela est dû au fait que les réseaux de confusion ne représentent pas toutes les hypothèses en compétition et qu'une partie de la "masse" totale des probabilités à distribuer parmi les mots est manquante. Les hypothèses en compétition ne sont pas toutes conservées à cause de la technique d'élagage nécessaire pour éviter de compromettre l'alignement final du réseau de confusion. Plus le graphe est grand, plus cette "masse" sera éparpillée. Pour lisser ces probabilités, une transformation linéaire par morceau est appliquée à la probabilité *a posteriori* d'un mot. Cette transformation est appelée mapping. Cette mesure associée au mapping de la probabilité *a posteriori* est appelée $m_{MAP}(w)$.

3 Combinaisons de mesures : références

Nous avons donc quatre mesures provenant de différentes parties du système de reconnaissance de la parole. Pour avoir le meilleur de chacune des mesures, plusieurs techniques de combinaison sont étudiées. Ainsi, plusieurs classificateurs et opérateurs d'aggrégation sont développés et leurs performances seront comparées à celles provenant d'aggrégations floues.

– **Moyenne arithmétique** Une des combinaisons utilisée comme référence est la moyenne arithmétique des quatre mesures.
– **Régression linéaire** Pour prendre en compte des qualités de chaque mesures, une simple interpolation linéaire peut être utilisée pour s'adapter à la modélisation prédictive de la justesse d'un mot comme dans l'équation :

$$y = \sum_{i=1}^{n} \alpha_i x_i + \alpha_0 \quad (4)$$

La méthode des moindres carrés est utilisée pour estimer les paramètres de l'interpolation sur un corpus spécifique.
– **Machines à Vecteurs de Support (SVM)** Les machines à vecteurs de support sont une des méthodes de classification étudiée ici.
– **Perceptron Multicouche (MLP)** Un perceptron multicouche avec rétropropagation et deux couches cachées est utilisé. Chaque mesure correspond à une entrée et il y a une unique sortie.

4 Techniques d'aggrégation floues

En général, l'incertitude en théorie des probabilités est vue en termes d'occurrences de fait connus. Dans une tâche de reconnaissance de la parole, ce qui est connu est le fait qu'une hypothèse soit correcte ou non. La probabilité est une notion intéressante quand il s'agit d'événements sériels qui requièrent une notion d'énumération de l'incertitude mais ne semble pas convenir pour introduire un degré d'accomplissement d'une situation connue [16]. En effet, quand on applique des mesures de confiance en reconnaissance de la parole, ce que l'on veut savoir est principalement le degré de justesse associé à une hypothèse de reconnaissance. Les systèmes d'inférence flous semblent pertinents pour cette tâche car ils proposent d'utiliser la notion d'incertitude du point de vue possibiliste.
Ces systèmes utilisent un ensemble de règles floues pour faire correspondre des entrées floues à des sorties floues. Ils permettent une classification de variables floues grâce à des règles en "si...alors" [21].
Plusieurs techniques utilisant la logique floue vont ici être développées afin d' en étudier la pertinence en reconnaissance de la parole.

4.1 Mesures de possibilité

Les mesures de confiances décrites dans le paragraphe 2 sont de nature probabilistes. Dans [6], les auteurs expriment la transformation entre probabilité et possibilité par : Soit C un ensemble d'événements composé de $c_1, c_2, ..., c_n$, soit $p_i = P(c_i)$ la mesure de probabilité associé à chacun des éléments de l'ensemble C où $p_1 \geq p_2 \geq ... \geq p_n$, et $\pi_i = \Pi(c_i)$ la mesure de possibilité associée à chacune des mesures de l'ensemble C, alors la solution optimale est :

$$\forall i = 1, n, \pi_i = \sum_{j=i}^{n} p_j \tag{5}$$

Cette transformation est alors utilisée pour obtenir les mesures de possibilité associées aux scores donnés par chacun des quatre experts. Parmi les techniques à notre disposition permettant de combiner ces scores possibilistes, nous utiliserons ici la fusion conjonctive normalisée, la fusion disjonctive et la fusion adaptative [5].

4.2 Intégrales floues

Dans cet article, deux intégrales floues sont utilisées, l'intégrale de Choquet et celle de Sugeno [27].

Soit $X = \{D_1, \ldots, D_n\}$ l'ensemble des n règles de décision (RD) et $\mathcal{P}$ l'ensemble des parties de X.
Définition 1 *Une* mesure floue *ou* capacité, μ, *définie sur* X *est une fonction* $\mu : \mathcal{P}(X) \longrightarrow [0, 1]$, *vérifiant les axiomes :*

$$\mu(\emptyset) = 0, \mu(X) = 1 \tag{6}$$

et

$$A \subseteq B \Longrightarrow \mu(A) \leqslant \mu(B) \tag{7}$$

Les mesures floues généralisent les mesures additives, en remplaçant l'axiome d'additivité par un axiome plus souple, celui de monotonie. Dans notre contexte de fusion de règles de décision, $\mu(A)$ représente l'importance, le degré de croyance dans une décision, apportée par un sous-ensemble A des RDs. Pour construire la décision finale, il s'agit de combiner les degrés de confiance partiel de chacune des RDs dans un degré de confiance globlal, en prenant en compte les différents poids de chacun.

L'intégrale de Choquet. L'intégrale de Choquet a été introduite en théorie des capacités [2, 24].

Definition 2 *Soit* μ *une mesure floue sur X. L'* intégrale de Choquet *sur des degrés de confiance* $\vec{\phi} = [\phi_1, \ldots, \phi_n]^t$ *notée* $\mathcal{C}_\mu(\vec{\phi})$*, est définie par :*

$$\mathcal{C}_\mu(\vec{\phi}) = \sum_{j=1}^{n} \phi_{(j)}[\mu(A_{(j)}) - \mu(A_{(j+1)})] \tag{8}$$

où $(.)$ *est une permutation telle que* $(i) \leq (j) \Rightarrow \phi(i) \leq \phi(j)$ *et* $A_{(j)} = \{(j), \ldots, (n)\}$ *représente les* $[j..n]$ *critères associés en ordre croissant et* $A_{(n+1)} = \emptyset$*.*

Déterminer la mesure floue
Il existe plusieurs méthodes pour déterminer la mesure floue la plus adéquate pour une application donnée [12]. Des algorithmes ont été développés pour intégrer au mieux les problèmes d'initialisation des treillis associés à la mesure floue (matrices mal conditionnées, convergence, temps de calcul). L'objectif est de trouver une approximation de la mesure floue en minimisant un critère d'erreur. À notre connaissance, l'algorithme fournissant la meilleure approximation et le mieux adapté à notre problème, est celui proposé par M. Grabisch [10]. Il part du principe qu'en l'absence d'information le modèle d'agrégation le plus raisonnable est la moyenne arithmétique. À partir d'un ensemble d'alternatives - ensemble d'échantillons caractéristiques et valeur de l'intégrale attendue - cet algorithme apprend la mesure floue en se fondant sur un principe de descente de gradient avec contraintes. L'idée est de minimiser l'erreur, au sens des moindres carrés, entre la valeur de l'intégrale calculée sur la mesure floue associée et la sortie attendue. Par ailleurs, cette approche permet de conserver une cohérence même si des données sont manquantes pour traiter tous les chemins du treillis. Enfin, l'apprentissage est très rapide. Cette aggrégation, utilisant l'intégrale de Choquet est notée m_{CI}.
Les mesures floues utilisées dans la méthode de Choquet peuvent également être obtenues grâce aux mesures de possibilité décrites dans le paragraphe 4.1. Cette aggrégation est notée m_{CP}.

L'intégrale de Sugeno. L'intégrale de Sugeno est décrite dans [11]. Les systèmes fondés sur l'intégrale de Sugeno ont déjà prouvé leurs bonnes performances en termes de

classification [14]. Un tel système a été utilisé dans [13] avec de bons résultats sur une base de données de parole en Espagnol. Il s'agit d'une base de données téléphonique échantillonées à 8kHz. Sur de la parole continue (environ 9405 mots), le taux d'égale erreur est passé de 22.85 avec un MLP à 22.05 avec un système fondé sur l'intégrale de Sugeno.

Déterminer la mesure floue
Pour les expérimentations, nous utilisons le même type d'intégrale de Sugeno que celle utilisée dans [13], obtenant ainsi 16 règles. La fonction d'appartenance gaussienne est utilisée pour "fuzzifier" les scores donnés par le système de reconnaissance. Cette méthode d'aggrégation utilisant l'intégrale de Sugeno est notée m_{SI}

Calcul d'indices. Dès que la mesure floue est entraînée, il est possible d'interpréter la contribution de chaque règle de décision dans la décision finale. Plusieurs indices peuvent être extraits à partir de la mesure floue pour mieux analyser le comportement et l'influence des règles de décision [11].

Indices d'importance
L'indice d'importance est fondé sur la définition proposée par Shapley dans le cadre de la théorie des jeux [26] et replacé dans le contexte des mesures floues par Murofushi et Soneda [23]. Son expression pour une mesure floue μ et une règle i est la suivante :

$$\sigma(\mu, i) = \frac{1}{n} \sum_{t=0}^{n-1} \frac{1}{\binom{n-1}{t}} \sum_{\substack{A \subseteq X \setminus \{i\} \\ |A|=t}} [\mu(A \cup \{i\}) - \mu(A)] \quad (9)$$

La valeur de Shapley peut être interprétée comme la valeur moyenne pondérée de la contribution $\mu(A \cup \{i\}) - \mu(A)$ de la règle de décision i parmi toutes les combinaisons. Une propriété intéressante est que la somme de toutes les valeurs relatives à toutes les règles de décision est égale à 1. En d'autres termes : $\sum_{i=1}^{n} \sigma(\mu, i) = 1$. Ainsi une règle de décision avec un degré d'importance plus petit que $1/n$ peut être interprétée comme une importance faible pour la décision finale.
La mesure utilisant les indices d'importance est notée m_{II}.

Indices d'interaction
L'indice d'interaction introduit par Murofushi et Soneda représente le degré d'interaction positif ou négatif entre deux règles de décision. Si la mesure floue a un comportement monotone alors des RDs interagissent. La valeur de l'interaction entre i et j, conditionnée par la présence des éléments de la combinaison $A \subseteq X \setminus \{ij\}$ est donnée par :

$$(\Delta_{ij}\mu)(A) = \mu(A \cup \{ij\}) + \mu(A) - \mu(A \cup \{i\}) - \mu(A \cup \{j\}) \quad (10)$$

En étendant ce critère sur tous les sous-ensembles de $A \subseteq X \setminus \{ij\}$ on obtient une évaluation de l'interaction entre les RDs i et j, comme suit :

$$I(\mu, ij) = \sum_{A \subseteq X \setminus \{ij\}} \frac{(n-t-2)!t!}{(n-1)!} (\Delta_{ij\mu})(A) \quad (11)$$

Une interaction positive pour deux RDs i et j signifie que l'importance d'une règle de décision est renforcée par la seconde. En d'autres termes, les deux RDs sont complémentaires et leur combinaison va en améliorant la décision finale. Une interaction négative signifie que les RDs sont antagonistes et que leur combinaison réduit l'impact de la décision finale.
Cet indice sera par la suite utilisé pour sélectionner les mesures utiles et complémentaires dans l'aggrégation.

5 Expériences

Les expériences ont été menées sur le corpus ESTER. ESTER est une campagne d'évaluation de systèmes de transcriptions d'émissions radiophoniques en français qui a démarré en 2003 et qui s'est terminé en janvier 2005 [9]. Le système de reconnaissance utilisé pour ces expériences est celui du LIUM (Laboratoire d'Informatique de l'Université du Maine) [4]. Ce système est basé sur le décodeur CMU Sphinx 3.3. Plusieurs éléments ont été rajoutés comme l'adaptation des modèles acoustiques utilisant la méthode SAT (Speaker Adaptative Training) ou encore ou encore le rescoring de graphes de mots utilisant des modèles de langage quadrigrammes. Le vocabulaire utilisé par le système du LIUM contient environ 65K mots. Les modèles acoustiques et de langage ont été appris sur le corpus d'apprentissage officiel d'ESTER. Le modèle de langage est un modèle quadrigramme obtenu par rescoring de graphe. Ce système a atteint la seconde position de la campagne avec 23.6% de taux d'erreur mot [4].

5.1 Apprentissage des paramètres pour les mesures de confiance

Les mesures de confiance ont été élaborées grâce à un échantillon de 4h provenant de France Inter, France Info, RTM et RFI. Ces 4h sont indépendantes du corpus d'apprentissage des modèles acoustiques et linguistique mais également du corpus nommé par la suite MTrain (voir sous-section suivante) et nous disposons de leur transcription manuelle. Le système de reconnaissance nous a permis d'en obtenir également une transcription automatique. À partir des transcriptions à la fois automatique et manuelle, nous avons pu calculer les différents paramètres de ces mesures. Pour la mesure de confiance acoustique, les résultats nous ont permis d'obtenir les paramètres μ, σ et a de l'équation (2). Pour la mesure de confiance LMBB, nous avons calculé les scores de confiance de la mesure LMBB à partir du taux d'erreur obtenu pour les mots appartenant aux différentes classes de LMBB.

5.2 Corpus utilisé pour l'évaluation des techniques d'aggrégation

Trois corpora ont été utilisés pour l'évaluation des différentes techniques d'aggrégation étudiées ici. Ces corpora proviennent du corpus de test officiel d'ESTER et consistent en 10h de parole continue. Chacun des corpus contient donc environ 3h20 de parole, ce qui équivaut à environ 31000 mots. Ils sont notés MTrain, MDev et MTest. MDev est utilisé en tant que corpus de validation pour les techniques de MLP et Sugeno notamment. Pour rester cohérents dans notre choix du corpus de test, nous avons pris les 3h20 du corpus qui ont été enregistrées le plus tardivement. Ceci a conduit à des enregistrements provenant de radios qui ne sont pas présentes dans MTrain et MDev. En résumé, MTrain est composé d'1h de Radio Classique, 1h de France Inter, 1h de France Info et 20 minutes de Radio France International (RFI). MDev est composé de 1h de France Culture, 1h de France Info, 1h de France Inter et 20 minutes de RFI. MTest est composé d'1h de RFI et 2h20 de Radio Televison Marocaine.

5.3 Métrique d'évaluation

La métrique utilisée ici pour montrer qu'une mesure de confiance est pertinente pour évaluer la justesse d'un mot est le Confidence Error Rate [28]. Elle correspond à :

$$CER = \frac{\text{Nombre d'étiquettes incorrectement assignées}}{\text{Nombre total d'étiquettes}} \tag{12}$$

La décision d'étiqueter un mot comme étant correct ou de l'étiqueter incorrect dépend d'un seuil qui est optimisé sur le corpus MTrain. Il doit être minimal pour le CER. Avec ce seuil, la mesure de confiance peut être évaluée sur MTest.

5.4 Mesures simples

Le tableau 5.4 décrit les résultats obtenus en terme de CER pour les quatre mesures de confiance de notre étude.

Mesure de confiance	MTrain (%)	MTest (%)
Taux de mots émis incorrects	17	22.3
WP	14.65	**17.88**
MAP	14.65	**17.88**
AC	16.04	22.18
LMBB	16.05	22.17

TABLE 1 – Résultats en termes de CER pour les quatre mesures de confiance

Les mesures les plus pertinentes sont les mesures WP et MAP qui obtiennent un score de 17.88% en termes de CER sur MTest. La technique de mapping préserve le score CER car une transformation linéaire par morceaux n'affecte pas la distribution des scores par rapport au seuil optimal. Le taux de mots émis incorrects permet d'avoir une référence pour le CER et montre que toutes les mesures prises séparemment sont pertinentes pour attester de la justesse d'un mot.

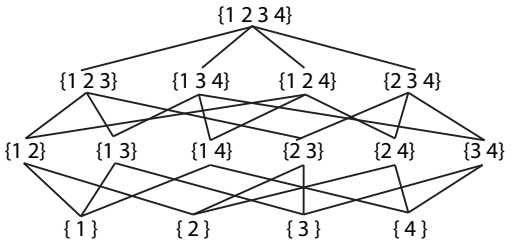

FIGURE 1 – Treillis pour quatre mesures/experts. Chaque chemin est associé à un poids calculé par la technique basée sur l'intégrale de Choquet. $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$ correspondant aux poids associés aux mesures WP, MAP, AC et LMBB.

5.5 Résultats après aggrégation

Plusieurs combinaisons des quatre mesures ont été développées pour cette étude. Quelques exemples de référence telles que la moyenne arithmétique, une interpolation linéaire ou encore un MLP sont mis en oeuvre pour comparer les techniques. Afin de prendre en compte la notion d'incertitude du point de vue possibiliste, plusieurs systèmes d'inférence floue sont utilisés pour aggréger les quatre mesures de confiance.
Par exemple, la figure 1 illustre comment la méthode basée sur l'intégrale de Choquet calcule les poids utilisés dans le processus de fusion (section 4). Ce treillis est une représentation de l'ensemble des parties de N (appelées capacités pour Choquet et mesures floues par Sugeno). Chaque chemin du treillis mène à une partie du plan où chaque mesure est associée à un poids p. En se basant sur l'ordre des scores des experts pour un mot donné, cela va entraîner des poids différents à attribuer à chacun des experts. Par exemple, si un mot w a des scores de confiance suivant l'équation :

$$m_{WP}(w) < m_{MAP}(w) < m_{AC}(w) < m_{LMBB}(w) \tag{13}$$

Le score final m_{CI} associé au mot w sera :

$$\begin{aligned} & m_{WP}(w) * (p(\{1234\}) - p(\{234\})) \\ +& m_{MAP}(w) * (p(\{234\}) - p(\{34\})) \\ +& m_{AC}(w) * (p(\{34\}) - p(\{4\})) \\ +& m_{LMBB}(w) * p(\{4\}) \end{aligned} \tag{14}$$

Le tableau 2 détaille les résultas obtenus pour chacune des techniques testées. Les fusions utilisant les mesures de possibilité ne paraissent pas pertinentes. Les SVM semblent surapprendre car elles obtiennent un bon score au niveau de l'apprentissage mais les données de test étant assez différentes des données d'apprentissage, le taux de CER chute. Chacune des autres techniques améliore les résultats d'une simple interpolation linéaire ou de la moyenne arithmétique et sont pertinentes pour attester de la justesse d'un mot. Comme le corpus d'apprentissage est conséquent, l'intégrale de Choquet n'améliore pas les résultats

Mesure aggrégée	MTrain (%)	MTest (%)
Moyenne arithmétique	14.18	17.61
Régression linéaire	13.92	17.55
SVM	13.83	22.16
MLP	13.87	**17.44**
Fusion conjonctive normalisée	16.04	22.16
Fusion disjonctive	16.04	22.15
Fusion adaptative	16.04	22.17
Sugeno m_{SI}	13.72	**17.44**
Choquet m_{CP}	14.72	21.16
Choquet m_{CI}	13.90	**17.49**

TABLE 2 – Résultats en termes de CER pour les différentes techniques d'aggrégation.

de la technique basée sur l'intégrale de Sugeno mais les deux techniques sont performantes en termes de CER.

6 Sélection des mesures

Pour estimer la corrélation entre les différentes mesures et mieux appréhender l'importance de chacune des mesures dans l'aggrégation, les indices d'importance et d'interaction sont calculés (voir paragraphe 4.2). Ces indices sont fournis par le treillis de Choquet et seront ensuite utilisés avec cette intégrale pour observer la pertinence de la sélection de mesures pour l'aggrégation finale.

L'indice d'importance, décrit en 4.2, représente la corrélation entre le score donné par une mesure et la justesse d'un mot. Tous les indices d'importance de chacune des mesures sont détaillés dans le tableau 3.

Mesure de confiance	Indice d'importance
WP	0.517512
MAP	1.930417
AC	0.536041
LMBB	1.016030

TABLE 3 – Indices d'importance des quatre mesures

Une possibilité pour utiliser cet indice consiste en une simple interpolation linéaire des quatre mesures avec comme poids leur indice d'importance. Cette aggrégation est notée m_{II}.
L'indice d'interaction également décrit en 4.2, est également calculé. Les résultats sont fournis dans le tableau 4.

	WP	MAP	AC	LMBB
WP	0.000000	-0.255104	0.041425	-0.776167
MAP	-0.255104	0.000000	-0.418659	-1.405465
AC	0.041425	-0.418659	0.000000	0.102432
LMBB	-0.776167	-1.405465	0.102432	0.000000

TABLE 4 – Indice d'interaction des quatre mesures

L'idée en sélectionnant les mesures qui vont être gardées dans l'aggrégation est de rejeter la mesure ayant le plus faible indice d'importance et dont l'indice d'interaction est plus bas que la moyenne des indices d'interaction. Ceci a permis d'obtenir le tableau 5.

Mesure aggrégée	MTrain (%)	MTest (%)
Choquet (4 mesures)	13.90	**17.49**
m_{II}	13.82	**17.34**
Choquet (3 mesures)	13.98	18.22
Choquet (2 mesures)	13.91	**17.20**

TABLE 5 – Résultats en termes de CER pour l'aggrégation en sélectionnant les mesures

Ces résultats prouvent la pertinence des indices d'importance. En les utilisant comme coefficients d'une interpolation linéaire, ils améliorent les résultats du MLP et des deux systèmes d'inférence flous basés sur les intégrales de Choquet et de Sugeno avec quatre mesures. En sélectionnant les mesures gardées dans l'aggrégation, les meilleurs résultats sont obtenus en conservant seulement deux mesures dans la technique basée sur l'intégrale de Choquet.

7 Discussion et conclusions

Cette étude présente comment l'utilisation de la logique floue est pertinente pour les mesures de confiance en reconnaissance de la parole. Pour des applications où l'aspect intéressant des mesures de confiance est leur capacité à classifier correctement les hypothèses, le taux associé de CER est important. Par exemple, cela peut être utile si l'on veut que la mesure de confiance permette de bien détecter les mots corrects d'une transcription automatique afin d'augmenter le corpus d'apprentissage des modèles acoustiques de manière non supervisée. Les techniques utilisant la logique floue se sont montrées efficaces en termes de CER et permettent d'aggréger efficacement les qualités de chacune des mesures. De plus, les intégrales floues et celle de Choquet en particulier peuvent apporter une interprétation sémantique au résultat et une mise en perspective du problème. La technique basée sur l'intégrale de Choquet, en sélectionnant seulement deux mesures, améliore les taux de CER de 4.10% en comparaison des résultats obtenus avec les mesures simples.

Références

[1] D. Charlet, G. Mercier, and D. Jouvet. On combining confidence measures for improved rejection of incorrect data. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Aalborg, Danemark, Septembre 2001.

[2] G. Choquet. Theory of capacities. In *Annales de l'Institut Fourier*, pages 131–295, 1953.

[3] A. Cornuéjols and L. Miclet. *Apprentissage artificiel : concepts et algorithmes*. Eyrolles, 2002.

[4] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. The LIUM speech transcription system : a CMU

Sphinx III-based system for french broadcast news. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, Septembre 2005.

[5] D. Dubois and H. Prade. Possibility theory in information fusion. *Data Fusion and Perception (CISM Courses and Lectures)*, 431 :53–76, 2001.

[6] D. Dubois, H. Prade, and S. Sandri. On possibility/probability transformations. In *Proc. of the Fourth IFSA Conference*, pages 103–112, Seoul, Korea, 1993.

[7] G. Evermann and P.C. Woodland. Large vocabulary decoding and confidence estimation using word posterior probabilities. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Juin 2000.

[8] G. Evermann and P.C. Woodland. Posterior probability decoding, confidence estimation and system combination. In *Speech Transcription Workshop*, 2000.

[9] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of french broadcast news. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, Septembre 2005.

[10] M. Grabisch. A new algorithm for identifying fuzzy measures and its application to pattern recognition. In *IEEE International Conference on Fuzzy Systems*, pages 145–150, 1995.

[11] M. Grabisch. The application of fuzzy integrals in multicriteria decision making. In *European Journal of Operational Research*, pages 445–456, 1996.

[12] M. Grabisch and Nicolas J.M. Classification by fuzzy integral - performance and tests. In *Fuzzy Sets and Systems, Special Issue on Pattern Recognition*, pages 255–271, 1994.

[13] G. Hernandez-Abrego and J.B. Marino. Fuzzy reasoning in confidence evaluation of speech recognition. *IEEE International Workshop on Inteligent Signal Processing WISP'99*, Septembre 1999.

[14] J.-S. R. Jang. ANFIS : Adaptative-network-based fuzzy inference system. *IEEE Transactions on systems, man and cybernetics*, 23(3) :665–685, Mai/Juin 1993.

[15] H. Jiang. Confidence measures for speech recognition : a survey. *Speech Communication Journal*, 45 :455–470, 2005.

[16] M. Laviolette and J.W. Seaman Jr. The efficacy of fuzzy representations of uncertainty. *IEEE Transactions on Fuzzy Systems*, 2(1) :4–15, Février 1994.

[17] H. Mangu, E. Brill, and Stolcke A. Finding consensus in speech recognition : Word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4) :373–400, 2000.

[18] J.-L. Marichal. Aggregation of interacting criteria by means of the discrete choquet integral. In *Physica-Verlag GmbH*, pages 224–244, 2002.

[19] J. Mauclair, Y. Estève, S. Petit-Renaud, and P. Deléglise. Automatic detection of well recognized words in automatic speech transcriptions. In *LREC, Language Resources and Evaluation*, Genoa, Italy, Mai 2006.

[20] J. Mauclair, L. Wendling, and D. Janiszek. Fuzzy integrals for the aggregation of confidence measures in speech recognition. In *Proc. of the IEEE International Conference on Fuzzy Systems*, pages 1149–1156, Taipei, Taiwan, Juin 2011.

[21] J.M. Mendel. Fuzzy logic systems for engineering : a tutorial. *Proceedings of the IEEE*, 83(3) :345–377, Mars 1995.

[22] F. Metze, T. Kemp, T. Schaaf, T. Schultz, and H. Soltau. Confidence measure based language identification. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Juin 2000.

[23] T. Murofushi and S. Soneda. Techniques for reading fuzzy measures(III) : interaction index. In *Proc. of the 9th Fuzzy System Symposium*, pages 693–696, Sapporo, Japan, Mai 1993.

[24] T. Murofushi and M. Sugeno. A theory of fuzzy measures : representations, the choquet integral, and null sets. In *Journal of Mathematical Analysis and Applications*, pages 532–549, 1991.

[25] R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, and J. Pardo. Confidence measures for spoken dialogue systems. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, Mai 2001.

[26] L. Shapley. A value for n-person games. In *Contributions to the Theory of Games, Annals of Mathematics Studies*, pages 307–317, 1953.

[27] M. Sugeno. Fuzzy measures and fuzzy integrals : a survey. *Fuzzy Automata and Decision Processes*, pages 89–102, 1977.

[28] F. Wessel, K. Macherey, and R. Schlüter. Using word probabilities as confidence measures. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 225–228, Seattle, USA, Mai 1998.

[29] F. Wessel and H. Ney. Unsupervised training of acoustic models for large vocabulary continous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13 :23–31, 2005.

[30] R. Zhang and A. Rudnicky. Word level confidence annotation using combinations of features. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, pages 2105–2108, Aalborg, Danemark, Septembre 2001.